sponsorship, and (d) the round-figure magnitude of the program (i.e., number of projects, participating school districts, sponsors, and so on).

2. Describe the evaluation design against the backdrop of program development.

3. Identify five issues that were reconciled one way, but might have been reconciled in other ways, and note how these decisions have influenced the evaluation of Follow Through.

In the discussion later, I hope there will be an opportunity to suggest some lessons that Follow Through has provided for planning and carrying out future social experiments in natural settings.

Follow Through Program

Congress authorized Follow Through in 1967 under an amendment to the Economic Opportunity Act to provide developmental and educational services for poor children in primary grades who had experienced Head Start or equivalent preschool. A large-scale service program, roughly similar in scope to Head Start, was envisioned originally and reflected in the enabling legislation. Appropriations were not sufficient, however, so the program was re-cast as an R&D program even though the language of the legislation itself was never changed. This R&D orientation has had complex effects on the way the program (and its evaluation) evolved and how various stakeholders view and judge the program.

Since Follow Through was authorized under the EOA, the Office of Economic Opportunity had responsibility for it just as they did for Head Start. OEO

or program priorities. Also, every individual project applies some additional variations on sponsor's objectives that will be sought locally. The measurement effort necessary to embrace the full sweep of all objectives held by all sponsors and all individual projects has never yielded gracefully to solution. As a consequence, it is inevitable that the measures obtained can be criticized with some justification by every sponsor or project as not including "all the things I am trying to do."

Overview of Evaluation Design

The basic design is that of a before-and-after experiment. Follow Through program participants define the experimental groups and nonparticipants of similar characteristics comprise the controls. "Before" (in the before-and-after design) refers to measures obtained as near as possible to the initiation of the experimental treatment. This has meant measures of pupil characteristics as close to the beginning of a child's entering school year as could be managed administratively. "Before" measures on other participants, such as parents and teachers, have usually been obtained at a later point in the first year of school.

The final "after" point cannot yet be described historically. During early planning in 1968-69, the intended meaning of "after" was a point as near the end of the child's Third Grade as could be managed. By a decision adopted in Spring 1970, children entering their first school year of Follow Through in Fall 1969 formally became the first cohort eligible for evaluation; children who entered in Fall 1968 were officially designated as the

January 15, 1973

CONFLICTING VIEWPOINTS THAT AFFECT DESIGN, ANALYSIS, INTERPRETATIONS, AND REPORTING IN THE NATIONAL FOLLOW THROUGH EVALUATION

Philip H. Sorensen
Stanford Research Institute

We're drawn together by common interests in early childhood education or evaluation or both. Probably each of us carries an axe he'd like to grind, or even swing.

I'd like to be a lens grinder instead of an axe grinder, at least in these opening remarks. Since my perspective on both early childhood education and the evaluation of early childhood programs has been so totally dominated by responsibilities in the Follow Through program since July 1968, the lens I'll try to grind is one through which we can look at the Follow Through program and its evaluation. I want to focus on a selected few issues that have influenced the evolution of both the Follow Through program and its evaluation. My purpose is to provide a perspective from which constructive assessment of the Follow Through evaluation can be made. (We can grind and swing our axes in the later give-and-take.)

My remarks cover the following:

1. Give a brief overview of Follow Through (on the assumption that not all those present are familiar with it). This overview will include (a) initiation of the program, (b) transition from a service to an R&D orientation, (c) the concept of program

sponsorship, and (d) the round-figure magnitude of the program (i.e., number of projects, participating school districts, sponsors, and so on).

2. Describe the evaluation design against the backdrop of program development.
3. Identify five issues that were reconciled one way, but might have been reconciled in other ways, and note how these decisions have influenced the evaluation of Follow Through.

In the discussion later, I hope there will be an opportunity to suggest some lessons that Follow Through has provided for planning and carrying out future social experiments in natural settings.

Follow Through Program

Congress authorized Follow Through in 1967 under an amendment to the Economic Opportunity Act to provide developmental and educational services for poor children in primary grades who had experienced Head Start or equivalent preschool. A large-scale service program, roughly similar in scope to Head Start, was envisioned originally and reflected in the enabling legislation. Appropriations were not sufficient, however, so the program was re-cast as an R&D program even though the language of the legislation itself was never changed. This R&D orientation has had complex effects on the way the program (and its evaluation) evolved and how various stakeholders view and judge the program.

Since Follow Through was authorized under the EOA, the Office of Economic Opportunity had responsibility for it just as they did for Head Start. OEO

transferred funds to the Office of Education for program administration, but OEO retained a vested interest in the program's design and management.

Follow Through began in the 1967-68 school year when 52 projects were initiated in 40 purposively selected school districts. All these first projects were of the type that have come to be called "self-sponsored;" that is, each local district conceived and began to implement its own concept of a "best" or most appropriate program for impoverished primary grade children through grade three. This notion of intentionally diverse approaches was labeled "planned variation."

During 1967-68, the Follow Through program office in OE refined the concept of planned variation to more nearly approximate a systematic experiment. Various recognized proponents of early education approaches were invited to serve as "model sponsors" to install and support their approaches in one or more project locations. This mode of sponsorship began with the 1968-69 school year. Local preferences for models or approaches were honored -- districts chose their preferred approach (within loose limits) following a sort of "courtship" with interested sponsors. Note well that principles of local autonomy and mutual preference determined the sponsor-to-district affiliation rather than some criteria of experimental design. At the start of the 1968-69 school year, there were 106 projects and 15 sponsors, counting self-sponsored and parent implemented projects as "sponsor" categories. About half of the projects initiated in 1967-68 affiliated with a sponsor and the remainder continued as self-sponsored projects.

Some projects were added in 1969-70, bringing the total number to 160. In addition, six new sponsors joined the effort. More sponsors and a few

more projects were added in 1970-71, bringing the total projects to 177 and the number of sponsors to 22. A few more projects and two more sponsors were added in 1971-72 to bring the project total to its current level of 180 and sponsor total to 24 (ignoring, for simplification, a sponsor who opted out in 1969-70 and a few projects that also elected to terminate in 1969-70). About 92,000 children are involved in Follow Through this year.

The following are points to remember about Follow Through as we appraise its evaluation:

1. The selection of districts, schools, teachers, and children to be involved in Follow Through was judgmental and guided by rules of individual district or child eligibility (e.g., districts in poor communities, children from families that were poor by OEO poverty guidelines). None of the choices (with the possible exception of some participant selection within a project) were random.
2. Control groups were not established simultaneously with the experimental groups (i.e., Follow Through participants) by random selection from pools of children eligible for participation. All control groups were established after-the-fact.
3. The mode by which sponsors affiliated with school districts was purposive in the extreme and led to substantial imbalance in characteristics across projects within sponsor categories and between sponsor groupings.
4. I've not mentioned it specifically yet, but you can appreciate that each sponsor entertains a somewhat different set of objectives

or program priorities. Also, every individual project applies some additional variations on sponsor's objectives that will be sought locally. The measurement effort necessary to embrace the full sweep of all objectives held by all sponsors and all individual projects has never yielded gracefully to solution. As a consequence, it is inevitable that the measures obtained can be criticized with some justification by every sponsor or project as not including "all the things I am trying to do."

Overview of Evaluation Design

The basic design is that of a before-and-after experiment. Follow Through program participants define the experimental groups and nonparticipants of similar characteristics comprise the controls. "Before" (in the before-and-after design) refers to measures obtained as near as possible to the initiation of the experimental treatment. This has meant measures of pupil characteristics as close to the beginning of a child's entering school year as could be managed administratively. "Before" measures on other participants, such as parents and teachers, have usually been obtained at a later point in the first year of school.

The final "after" point cannot yet be described historically. During early planning in 1968-69, the intended meaning of "after" was a point as near the end of the child's Third Grade as could be managed. By a decision adopted in Spring 1970, children entering their first school year of Follow Through in Fall 1969 formally became the first cohort eligible for evaluation; children who entered in Fall 1968 were officially designated as the

"implementation cohort" and thereafter excluded from the evaluation. By this convention, the first genuine "graduates" of Follow Through are children who began Follow Through as First Graders in Fall 1969 and completed Third Grade in Spring 1972. Children who began Follow Through as Kindergartners in Fall 1969 will not complete Third Grade until Spring 1973. Current expectations are that four successive cohorts will be followed longitudinally. The fourth cohort entering Kindergarten in Fall 1972 will complete Grade Three in Spring 1976.

Major classificatory variables for analyses include (1) cohort (one through four), (2) grade stream (enter kindergarten or first grade), and (3) treatment (Follow Through vs. non-Follow Through within sponsor or approach categories and approach vs. approach according to differences between Follow Through and non-Follow Through). A variety of other independent variables are employed in the analyses, but it is not appropriate to discuss the detail of analyses or findings to date in this symposium.

Two additional design considerations should be noted, for they are critical to analyses, interpretations of findings and suggest lessons for future planning.

One issue is the frequency and coverage of measures intermediate between the "before" or entrance measures and the "after" or exit measures. Intermediate level measures have been obtained for a subset of projects on which entrance measures were taken so that repeated measures would be available on children in some projects. The number of projects in which intervening time measures are available is small relative to the number of projects with baseline measures. The decision to limit the number of projects to be included

in data collection at intermediate points for any one cohort of pupils was partly a matter of economy (and deference to the tolerance of individual projects for frequent across-the-board measurement). More significantly, this limitation provides an effective restraint upon the temptation to draw conclusions about the efficacy of an approach before two conditions can be satisfied: (1) completion of a full term of three or four years for a cohort group and (2) accumulation of a reasonable number of examples (i.e., projects) representing an approach.

The second issue is that of control groups. I noted that these were established after-the-fact for each project and not randomly assigned when experimental groups were created. This condition constrains statistical inference. In addition, these non-Follow Through controls often are quite dissimilar from the Follow Through samples on many key demographic and socio-economic indices. I will dismiss, as messy but manageable, the administrative and diplomatic problems of finding and inducing schools to collaborate as controls.

I hope you can see, from these comments, that Follow Through should not be viewed as an experiment in the classical sense. From the evaluator's viewpoint, it is a quasi- or even a pseudo-experiment in which there is literally no manipulative control of treatments. It's social experimentation on a fairly ambitious scale with all the blooming, buzzing confusion that characterizes efforts to be systematic in a natural setting.

Some Key Issues With Alternate Resolutions

The Service Orientation vs. the Research Orientation

A service orientation toward Follow Through would dictate assuring that only the most needy were participants in the program. In contrast, an experimental orientation would either assure equal neediness for participants and non-participants or would permit variation on the impoverishment scale so that a "treatment by poverty" interaction might be identified. A service orientation also would argue for a standard presentation of services to all who qualify, whereas an experimental orientation would encourage a greater variety of services. Finally, evaluating a program under a service orientation would require some pre-defined standards against which program success could be measured. In contrast, an experimental orientation would more likely ask whether inter-treatment differences existed, and if so, where, to what degree, and so on.

Officially, Follow Through is an R&D program. Many people appraise it, however, in a service program context. In addition, many of the operational decisions, especially at local project levels, have been made as though the program were clearly and certainly a compensatory service program. Most of these latter decisions have been ones that juggled and reassigned participants according to judgments about the needs of individual children.

Policy vs. Theory Orientation

The grossest possible question that the policy maker might ask is "Does it work?" The theoretician is more likely to be concerned with the conditions under which particular effects are observed; i.e., the how, why, and when questions. An analogical distinction between the two orientations can be

made by referring to the input-black box-output model familiar to operations research. A policy orientation would suggest contrasting input levels with output levels without necessarily addressing questions of what happens inside the black box. The theory orientation, while not disinterested in input-output differences, is particularly concerned with the processes inside the black box. A policy orientation also is more likely to ask cost effectiveness and cost benefit questions of the data, whereas the theory orientation may, in many cases, ignore the cost variable.

The importance of this policy vs. theory or research orientation is no more sharply felt than when one speaks to a mixed audience of policy makers or administrators on the one hand and, on the other, researchers who address questions of theories of instruction or human development. The "grand plan" for the evaluation was one that anticipated patience by policy makers and legislators while some how and why questions were pursued. We have experienced increasing pressure, however, to draw policy-type conclusions on the basis of incomplete evidence. They ask serious and legitimate questions that deserve sober answers -- such as, "Should Follow Through Approach X be continued?" -- but the evidence is partial, often inconsistent, and sometimes appears trivial. Policy makers show occasional impatience with answers like "just wait four more years and spend another umpteen dollars and we ought to be able to make a better estimate."

Formative vs. Summative Orientation

If one gives prominence to formative assessment, he is by that choice encouraging the program to change as it grows in response to frequent and

fairly rapid feedback. Summative assessment, on the other hand, is more congenial to a stable treatment observed over a sufficient period to permit conclusions to be drawn about the whole program or the relative strength of fixed alternatives. These comments are not meant to imply that one view of assessment is "better" than another but simply to note that formative assessment is most appropriate for those conditions that obtain when designing a system is the primary objective, and summative assessment is most appropriate to conditions where the objective is to test the worth of a describable system.

In the first year or two of the evaluation, it was felt that both orientations could be maintained simultaneously -- that a single agency could provide rapid turnaround of data to guide program design decisions and also manage the detachment and objectivity required for summative assessment. The load was more than we could manage, not due so much to data processing and turnaround time as with the differences in the kinds of data appropriate for the differing classes of questions. There's a fundamental incompatibility between emergent treatments, influenced by formative assessments, and experimental treatments that maintain consistency long enough to warrant inferences being drawn about their effects.

General vs. Specific Criteria of Success

This dichotomy might have been referred to better as abstract vs. concrete, or less measurable vs. more measurable outcomes. I don't know how to state a clear preference with regard to broad vs. narrow criteria or aggregated and gross outcome measures vs. disaggregated and fine-grain

measures. In a sense, it is a no-win game. Some of the approaches in Follow Through have never managed to state their objectives in more than very broad terms, thus creating difficult, if not insoluble, problems of measurement. On the other hand, those that have stated specific objectives in explicit terms (which lend themselves better to measurement) may be overemphasizing educationally trivial things. An objective of "increasing the life chances of poor children" is both noble and general, abstract, nonspecific, and unmeasurable. An objective of "teaching children to count dots," while defensible as an instrumental objective building toward competence in concepts of number (and therefore maybe contributing to future life chances) is a rather narrow criterion for judging a program. The golden mean is somewhere in between.

Frequent Reporting vs. Deferred Reporting.

This issue is related to an earlier one of formative vs. summative assessment and also to the policy vs. theoretical orientation. Certainly the confidence that one may express about observed findings is a function of the reliability of the measures and reliability is greater with repeated measures over an extended period. In Follow Through, the heart of the issue on frequent reports vs. deferred reports is whether all of the Follow Through approaches are going to be given a full period during which each of several cohorts can display their effectiveness. If comparative findings are reported frequently, then the risk that someone will draw premature conclusions is increased. On the other hand, if reports are deferred too long or qualified excessively when issued, their utility for the policy maker may be lost entirely.

Perhaps the balance can be struck by acknowledging the legitimacy of two distinguishable classes of reports. One would emphasize timeliness and best-estimate trends to reduce the chances of gross errors of either omission or commission by policy makers. The other kind of report would contain more definitive analyses that sought to explain phenomena, perhaps only partly perceived in reports where timeliness overrode precision in importance.